"EMPLOYEE INTERNET MANAGEMENT DEVICE"

RELATED APPLICATIONS

The present application claims priority under 35 U.S.C. § 119(e) from U.S. Provisional Patent Application Serial No. 60/175,937, filed January 12, 2000, which provisional application is incorporated herein by reference in its entirety.

FIELD OF THE INVENTION

The present invention relates to the field of employee Internet and network management. More specifically, it relates to a device and method for employers to monitor and maintain employee compliance with an organization's acceptable use policy relating to network usage.

BACKGROUND OF THE INVENTION

As more and more businesses connect to the Internet, employee misuse of company computers and networks is increasing and, perhaps, reaching epidemic proportions. As a result, employee productivity is often significantly reduced. Moreover, the threat of lawsuits against the company is increased due to increased potential for inappropriate or illegal use of the network.

Unfortunately, there are very few products available for use by managers to monitor network use by employees, whether the network is a local area network or a wide area network such as the Internet, and report on violations of corporate policy. Instead, rather than allow monitoring of network use, most products attempt to block access to web sites

that are deemed non-business-related sites. However, such products are largely ineffective, regularly allowing access to non-business-related sites, as well as erroneously blocking access to legitimate business-related sites. Most of these products compile databases of web uniform resource locators (URLs) that are deemed inappropriate. There are many problems with this approach. First, it addresses only web access, ignoring email, chat sessions, and similar communications. Second, the Internet is growing too rapidly to maintain an effective database of inappropriate sites. As soon as a new database update is released, it is already hopelessly out of date. Third, the size of the database must be proportional to the size of the Internet. Given the Internet's rapid and unlimited growth, no database approach can scale well enough to use in the long term. Fourth, the selection of appropriate versus inappropriate URLs is made by the manufacturer of the product. This reduces the manager's ability to tailor the database to reflect individual corporate needs.

Alternatively, a few products use lists of keywords rather than a URL database to monitor employee activity. These have the advantage of scaling well and enabling managers to customize web access rules that more accurately reflect corporate policy. However, most products use a simplistic implementation of keyword searching, resulting in nearly as many errors as with the URL database approach. For example, a legitimate medical site may be incorrectly identified as pornographic because of references to human genitalia.

SUMMARY OF THE INVENTION

The present invention utilizes a method of weighted regular expressions to perform language analysis, categorize the monitored data and report deviations from a company's

acceptable use policy. The present invention monitors all Transport Control Protocol/Internet Protocol (TCP/IP) network communications. It is not limited to just web or email monitoring. It stores any TCP/IP sessions that match the criteria selected by the user from either predefined categories or user defined keywords. The stored sessions can then be viewed, downloaded, and/or deleted by the user.

The search criteria are selected in two ways, by subject category or by keyword matching. Categories are pre-defined topics, such as "conflict," "resignation," or "shopping." The user can select whether the category should be on or off. If on, a sensitivity is selected by the user. Sensitivity levels are inversely proportional to the amount of category-related language required to indicate a match. For example, a low sensitivity requires more category-related language than a high sensitivity to qualify as a match. In addition, some categories are hierarchical, containing no regular expressions but depend upon matches by constituent categories. For instance, a "disgruntled" hierarchal category would generate a match if there were enough matches in its constituent categories, such as "resignation" and "conflict." A further hierarchal category, such as "work place violence," could generate a match if matches are generated in the "disgruntled" category and in a "weapons" category. The keywords are user-defined. The user can select whether any of the keywords or all of the keywords are required for a match. This is similar to the simple keyword matching used as the foundation of most keyword-based network monitoring products.

BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 is a flow diagram that describes how unprocessed logs or data, compiled by the present invention, are initially analyzed.

Fig. 2 is a flow diagram that describes how the data is further processed to determine if it meets selected search criteria.

BEST MODE FOR CARRYING OUT THE INVENTION

The hardware for implementing the present invention consists of a PC-compatible embedded-system that boots from flash ram and uses a hard disk for storing monitored data. A Linux operating system kernel and associated utilities, as well as proprietary software, is stored on the flash ram. The proprietary software allows for monitoring and storing raw TCP/IP data, for searching the raw data for the user-selected criterion and for providing a user interface via an integrated web server.

During monitoring and storing, the program listens to the Ethernet interface in promiscuous mode, storing each TCP/IP half-session to its own file or log on disk.

Next, the stored logs are processed by a separate program. Fig. 1 is a flow diagram which illustrates this processing. More specifically, it is desirable for the logs to be processed in roughly first in first out (FIFO) order. Therefore, the process always tries to read the oldest currently available log. If no logs are present, the process waits briefly, and tries again.

When a log has been successfully read, it is first examined to determine if it conforms to a known protocol. Unknown protocols will also be processed. However, the analysis and

reporting process can be enhanced for certain types of data, if the protocol is known. This protocol identification is accomplished by comparing the log data with known protocol patterns. If no pattern is discernible, the source and destination tcp port numbers are examined. If a well-known port number under 1024 is used, the data is assumed to be of the protocol associated with the tcp port. For example, the simple mail transport protocol (SMTP) can usually be identified by patterns within the data stream. However, if part of the data is missing, it may no longer be identifiable in this manner. But the number of either the source or destination tcp port will be 25. Because SMTP always uses tcp port 25, the log data can be assumed to be part of an SMTP session.

Some protocols transfer multiple independent data streams within a single tcp session. For example the SMTP protocol supports the transmission of multiple unrelated emails in one tcp session. If the log data is identified as being associated with one of these protocols, each independent part of the log is processed individually. Even these independent portions of the log may need to be broken down further into smaller pieces. For instance, email may contain multiple documents in the form of attachments, which may need to be converted, where possible, to a format containing text, and analyzed separately.

Each log, or independent portion of the log is then processed by the "categorize" subroutine which is illustrated in the flow diagram of Fig. 2. First, the data is stripped of any content which does not appear to contain language elements. The remainder, i.e., text containing language elements, is stored as a string of language elements separated by spaces. This allows the language elements or text to be effectively searched regardless of its original formatting. In example 1, addressed below, an email message is processed. Note that in

email, quotation of prior email references are commonly preceded by numerous "greater than signs" (>), which are stripped in this step.

The text is then searched to determine whether it matches the current set of user-selected criteria. If so, the log is saved in a separate file system in one or more subdirectories based on which criteria were matched. Then the log is deleted.

The criteria matching process is based on weighted key phrases or regular expressions. All key phrases take advantage of the "regular expressions" used in common Unix utilities such as egrep, sed, and perl. This enables the use of extremely flexible and powerful key phrases. Each category is assigned a numeric value. Each key phrase or regular expression within a category is also assigned a numeric value. When a log is examined, the sum of all values associated with each matching key phrase or regular expression is compared with the value for the category. If the sum meets or exceeds the category value, the file is considered a match for that category.

This process is different from simple keyword matching in that many individual key phrases can be matched, without necessarily causing a match for the category. It also enables matching based on a sufficient amount of questionable language content, the constituent key phrases of which might be completely innocuous individually.

Within each category, a regular expression can be assigned a positive or negative value. Using negative values facilitates avoidance of "false hits", or undesired matches. For example, in the medical web site example noted above, a legitimate web site would not necessarily produce a match for pornography if medical terms were assigned negative values and included in the key phrases within the pornography category. As another example, often

web-based news reports will contain language related to sports. Assume that a company wants to log sports-related activities, but doesn't want to log common news reports. This can be accomplished by assigning negative values to news-related key phrases and including these in the key phrase lists within sports. Much more sports language would then be required to trigger a match within a log containing news reports. This technique can be applied to any content that regularly produces false hits, effectively reducing a category's sensitivity level automatically whenever appropriate.

For example, the following is a category definition for a category relating to mergers and acquisitions:

```
acquisition (threshold = 4)

# resignation/recruiting
-4 resume (attached|enclosed)
-4 \b(his|her|your|my|a|the|attached|enclosed) resume
-4 resume[^ ]*\.(doc|rtf|html)

# News
-4 (top|front page|headline) (news|stor)
-4 today\'s headlines
-4 \(reuters\)
-4 \(ap\)
-4 \(upi\)
-2 edition
-2 \bnewsletter\b
-2 \bnews\b
-1 weekly

4 (buy|sell) (\w+ |) company
4 buyout
3 due diligence
2 stock (trade|shares)
2 merger
2 equity
2 \bipo\b
```

2 stock option
1 \bacqui[rs]
1 contract(\|s\|ed\|ing)\b
1 synergy

This category will find matches on merger/acquisition related activity. Note that any resignation or news related language will reduce the sensitivity, requiring additional merger/acquisition language to trigger a match. The log is searched for the weighted regular expressions in the order defined by the category definition. Thus, with respect to the acquisition category, the regular expression "resume (attached/enclosed)" will be searched first and the remainder of the weighted regular expressions will be searched in the order shown. Once a category's threshold value is met or exceeded, in this case 4, the search is stopped and the log is saved.

The following two examples show how logs are processed utilizing the categorize subroutine depicted in Fig. 2. The first example involves e-mail communications involving at least one employee and the second example relates to a web page located by an employee search.

- Example 1 - An email

MAIL From:<johndoe@company-a.net> SIZE=4414
RCPT To:<janedoe@company-b.com>
DATA
Received: from xyz.com ([10.74.91.90])
by some.mailserver.net (InterMail v03.02.03 118 118 102)
with ESMTP
id <19980924014702.FEMX9555@xyz.com>
for <janedoe@company-b.com>; Thu, 24 Sep 1998 01:47:02 +0000
From: "John Doe" <johndoe@company-a.net>
To: "Jane Doe" <janedoe@company-b.com>
Subject: Re: Harry's Resume
Date: Wed, 23 Sep 1998 19:46:24 -0600

X-MSMail-Priority: Normal
X-Priority: 3
X-Mailer: Microsoft Internet Mail 4.70.1161
MIME-Version: 1.0
Content-Type: text/plain; charset=ISO-8859-1
Content-Transfer-Encoding: 8bit
Message-Id: <19980924014702.FEMX9555@blah.blah.blah.com>

Jane,

The stock options would become fully vested upon any corporate merger or acquisition.

John,

>> John,
>>
>> The 5 year vesting period seems a bit long. Harry was wondering
>> what, if anything, would happen upon change of control.
>>
>> Jane,

>>> Jane,
>>>
>>> Harry's resume looks fine. I'll pass it on to the VP of Sales.
>>>
>>> Thanks,
>>> John

>>>> John,
>>>>
>>>> I think I found a good candidate for you. The attached
>>>> resume is from Harry Smith. He's interested in the
>>>> Colorado Sales Director position.
>>>>
>>>> Jane

Using the category definition defined above, the first match is on the text "attached resume" by the regular expression "\b(his|her|your|my|a|the|attached|enclosed) resume". Note that the email actually contains the lines:

>>>> I think I found a good candidate for you. The attached
>>>> resume is from Harry Smith. He's interested in the

The first thing that the categorize subroutine does (refer to flow chart) is to extract only language elements, therefore the common email quote characters, ">>>>", are not part of the text line that is searched. This match results in the sum being set to -4 because the term "resume" is assigned a weight of -4. Because the sum is not greater than or equal to 4 (the value for this category), the search continues. The next match is on the text "merger" by the regular expression "merger". The sum is updated to -2 because the term "merger" is assigned a weight of +2. Because sum is not greater than or equal to 4, the search continues. The next match is on the text "stock options" by the regular expression "stock option". The sum is updated to zero because the term "stock option" is assigned a weight of +2. Because the sum is not greater than or equal to 4, the search continues. The next match is on the text "acquisition" by the regular expression "\bacqui[rs]". The sum is updated to 1 because the weight assigned to this term is +1. Because sum is not greater than or equal to 4, the search continues. But, there are no more matches. Therefore, the text is not considered a match for this category and the log is deleted.

- Example 2 - A web page

```
<BASE HREF="http://www.mnanews.com/daily/">
<HTML>
<HEAD>
 <TITLE>
 Smart Buyers' News: Tech Stocks Fueling Merger, Acquisition Activity
 </TITLE>
</HEAD>
```

```
<body background="http://img.mnanews.com/images2/ebnback3.gif"
bgcolor="#FFFFFF" link="#0000FF" alink="#0000FF" vlink="#0000FF"
text="#XXXXXX" rightmargin="0" leftmargin="0" topmargin="0"
marginheight="0" marginwidth="0">

<!-- TITLE
-->
<!-- MAIN CONTENT AREA: INSERT ONE BLOCK OR A TWO-COL NESTED
TABLE -->
<!-- SITE LOGO -->
<img width="350" src="http://stuff.com/images/logo.gif" height="93" alt="Smart
Buyers' News">
<!-- STORY GOES HERE -->
<h2>
 Tech Stocks Fueling Merger, Acquisition Activity
</h2>
(7:00 p.m. EST, 1/21/98)
<br>
<i>By
<a href="mailto:johndoe@mycorp.com">John Doe
</a></i>
<p>The stock market's ongoing volatility is a key ingredient fueling the surging
merger and acquisition activity among technology companies, according to a report
released Wednesday by the New York investment firm Jane Doe & Associates.

<p>The total number of merger/acquisition transactions in the information
technology, media and communications industries climbed to new levels last year,
increasing 25% globally and 31% in North America, according to Bro
adview Associates' 1997 Technology M&A Report. Worldwide number of merger and
acquisition (M&A) transactions in the technology industry reached a record 4,040
in 1997, 25% more than the 3,224 transactions completed in 1996, according to
Broadview, a leading mergers and acquisitions investment bank serving the IT,
media and communications industries.

<p>-- The number of corporate buyouts jumped 57% in 1997; in the software sector,
the number of public sellers leaped 83%.

</BODY>
</HTML>
```

The first match is on the text " News:" by the regular expression "\bnews\b". The sum is set to -2 because the term has an assigned weight of -2. Because the sum is not

greater than or equal to 4, the search continues. The next match is on the text "buyouts" by the regular expression "buyout". The sum is updated to 2 because the term has an assigned weight of +4. Because the sum is not greater than or equal to 4, the search continues. The next match is on the test "Merger" by the regular expression "merger." The sum is updated to 4 because this term has an assigned weight of +2. Because sum is greater than or equal to 4, the log is saved and the search is finished for this category.

Through the use of common gateway interface (CGI) scripts, all reporting and maintenance is accomplished via a Web interface. To enhance ease-of-use, reports come in a number of different formats including, for example, reports based upon an individual employee or address and further showing, in bar graph format, the number of matches in each category. In addition, all bar graph segments are html links to more detailed reports, enabling the user to "drill down" via a graphical web interface. Moreover, the log viewer CGI script is capable of presenting a variety of data formats in an easy-to-read format to the user. For instance, a logged web page will be shown to the user as a web page, rather than as raw html. A logged binary print file in the HP PCL format will be displayed as text, rather than an illegible jumble of characters. Thus, the user will find stored data easy to understand.

The foregoing description of the invention relates to the best mode of practicing the invention known to the inventor at the time of filing this application. Alternatives will likely be recognized by those skilled in the art following a review of this patent. Such modifications or alternative approaches, recognized by those skilled in the art, are deemed to be part of this disclosure and within the scope of the present invention.